UNCLASSIFIED

AD NUMBER
ADB008303
LIMITATION CHANGES
TO: **Approved for public release; distribution is unlimited.**
FROM: **Distribution authorized to U.S. Gov't. agencies only; Test and Evaluation; NOV 1975. Other requests shall be referred to Rome Air Development Center, Griffiss AFB, NY 13441.**
AUTHORITY
RADC ltr, 12 Apr 1978

THIS PAGE IS UNCLASSIFIED

2

RADC-TR-75-287
Final Technical Report
November 1975

AD B008303

AD No. ____
DDC FILE COPY

ALPHA/NUMERIC EXTRACTION TECHNIQUE

Threshold Technology Inc., Delran, N.J.

S/C 390798

Distribution limited to U. S. Gov't agencies only; test and evaluation; November 1975. Other requests for this document must be referred to RADC (IRAP), Griffiss AFB NY 13441.

Rome Air Development Center
Air Force Systems Command
Griffiss Air Force Base, New York 13441

This report has been reviewed and approved for publication.

APPROVED:

BRUNO BEEK
Project Engineer

APPROVED:

HOWARD DAVIS
Technical Director
Intelligence & Reconnaissance Division

ACCESSION for
NTIS White Section
DDC Buff Section
UNANNOUNCED
JUSTIFICATION
BY
DISTRIBUTION/AVAILABILITY CODES
Dist. AVAIL. and/or SPECIAL
B

FOR THE COMMANDER:

JOHN P. HUSS
Acting Chief, Plans Office

Do not return this copy. Retain or destroy.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

18 19

1. REPORT NUMBER: RADC-TR-75-287
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): ALPHA/NUMERIC EXTRACTION TECHNIQUE. (6)
5. TYPE OF REPORT & PERIOD COVERED: Final Technical Report, (9)
6. PERFORMING ORG. REPORT NUMBER: N/A
7. AUTHOR(s): Phillips B. Scott (10)
8. CONTRACT OR GRANT NUMBER(s): F30602-75-C-0238 (15)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Threshold Technology Inc, 1829 Underwood Blvd, Delran NJ 08075
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 64701B, 43020304
11. CONTROLLING OFFICE NAME AND ADDRESS: Rome Air Development Center (IRAP), Griffiss AFB NY 13441
12. REPORT DATE: November 1975 (11) (12) 34p.
13. NUMBER OF PAGES: 26
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Same (16) AF-4302 (17) 430203
15. SECURITY CLASS. (of this report): UNCLASSIFIED
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A
16. DISTRIBUTION STATEMENT (of this Report): Distribution limited to U. S. Gov't agencies only; test and evaluation; November 1975. Other requests for this document must be referred to RADC (IRAP), Griffiss AFB NY 13441.
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report): Same
18. SUPPLEMENTARY NOTES: RADC Project Engineer: Bruno Beek (IRAP)
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): Speech Recognition; Pattern Recognition; Acoustic Phonetics
20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

This report describes the design, principles of operation and performance characteristics of an Advanced Development Model of a limited-vocabulary word recognition system. This system is intended to be used as an interface between a cartographic operator and a digitizing computer to allow voice entry of bathometric depth data from smooth sheets. The system is intended to accommodate a 15-word vocabulary, typically 10 digits and five control words. The system can store voice characteristics for six to eight operators. By the use of a numeric display located on the digitizing table cursor, OVER

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

390798 ✓ 1/B

Cont'd

the operator using the system can verify that each data digit spoken into the system has been correctly recognized. Data can then either be corrected, or if correct, entered into the digitizing system by the use of spoken control words.

This cartographic word recognition system is based upon the Threshold Technology, Inc. (TTI) commercial VIP-100 isolated-word recognition system. The VIP-100 can be automatically adapted on-line for individual speakers and/or words. The principal speech recognition modules of the VIP-100 are a speech preprocessor designed and manufactured by TTI and a general purpose minicomputer running with TTI designed software. For this contract, RADC furnished as GFE, a Data General Nova 1200 minicomputer with 8K memory to be included in the system.

To confirm system performance, accuracy tests were conducted from tape recordings. Each of 20 talkers recorded 360 test words and 150 training words. The training word sets consisted of 10 repetitions of each digit and each of the five control words. The test word sets consisted of 24 subsets of the complete vocabulary of 15 words. Recognition accuracy for the 20 talker set was in excess of 99% when the system was tested with this tape recorded data.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Section I

INTRODUCTION

The objective of this program has been to provide an automatic speech recognition (ASR) system suitable for use by cartographers for entering bathometric readings from smooth sheets into a digitizing computer. The ASR system will allow the cartographer to simultaneously obtain X-Y coordinate locations and provide voice data entry of depth readings for each coordinate. With his hands free to concentrate on the X-Y positioning device, the operator can speak the bathometric numbers and if they are correctly recognized, he can enter them by voice command into the computer without losing sight of the smooth sheet. Presently, he is diverted by the requirement to enter the bathometric readings by keyboard. This breaks his concentration on the visual and manual processes required. With the ability to enter the bathometric data by voice, a cartographer will be able to maintain both visual and manual concentration on the smooth sheet from which the data is being obtained.

The cartographic ASR system is a version of the basic off-the-shelf VIP-100 limited-vocabulary isolated-word ASR system supplied by Threshold Technology for many applications in both industry and Government. The Nova 1200 computer included in the system was supplied GFE by the Government. The VIP-100 in this application has been configured with 2400 baud Teletype interface for connection to the bathometric digitizing computer. Through this connection, the VIP-100 functionally is a direct replacement for a keyboard as an input to the bathometric computer. ASCII characters representing the word recognition decisions of the VIP-100 are transferred from the VIP-100 via the Teletype link to the bathometric computer. The system has the capability of recognizing a vocabulary of ten digits and five control words. Reference data for six operators can be stored in the system at any one time.

By mutual agreement between the contractors who have supplied the bathometric digitizing system and TTI, the cursor mounted operator's display is under control of the bathometric digitizing computer and will be supplied by the contractor for that system. An auxiliary display to be used principally for optimizing the ASR system for each operator's voice training has been supplied as a module of the VIP-100 system.

Before an operator uses the VIP-100 in the recognition mode, the system is first optimized for both the vocabulary words selected and for the operator's particular manner of speech pronuniciation by the use of a training routine. The operator speaks several utterances of each word in the vocabulary during the training operation. After training has been completed, the system will be ready to recognize the chosen vocabulary words when they are spoken by the operator that trained the system. It is not necessary to retrain the system each time a particular operator uses the system. The operator training data may be stored in the active computer memory or on punched paper tape for use when needed. The appropriate paper tape with the stored data can be read into the system whenever the operator or vocabulary is changed. The system may be retrained for a single word, multiple words, or the complete vocabulary at any time in order to accommodate vocabulary word substitutions or temporary changes in an operator's speech characteristics which may result from colds or other respiratory ailments.

To operate the system, the operator wears a lightweight, boom-mounted, noise-cancelling microphone and enters the spoken commands into the system through the Voice Input Remote Control Unit located conveniently to the operator. The Voice Level meter on this unit indicates to the operator that he is pronouncing the words at his normal intensity and aids in speaking level control adjustments.

System tests involving 20 talkers uttering 360 words each in addition to training samples were conducted. Results of these tests showed accuracy of approximately 99.4%. A complete description of the VIP-100 system as modified for this application is included in Section II of this report. System tests with results are described in Section III. Conclusions and recommendations are listed in Section IV.

Section II

TECHNICAL DISCUSSION

A. Introduction

To most expeditiously design and construct an Advanced Development Model of a highly reliable limited-vocabulary word-recognition system for this application, Threshold Technology Inc. (TTI) has supplied its commercial limited-vocabulary, isolated-word recognition system, the VIP-100. The principal hardware modifications were the substitution of a simplified numeric display (to be used principally for optimizing the system for each operator's voice) in place of an alphanumeric display usually included in a standard system, and a special interface, compatible with the RADC cartographic digitizing computer. Custom software was constructed as necessary to provide the necessary system functions.

The VIP-100 version which was supplied for this program includes as principal components a speech preprocessor, and a Nova 1200 minicomputer manufactured by Data General Corporation with 8K of core memory. The Nova 1200 was supplied as GFE by the Government. A Teletype Model ASR 33 is used for control and data input/output functions. Two Telex Model 1200 noise-cancelling microphones are used for speech input to the system. A custom display has been provided principally for use during training of the system. Training is the process of optimizing the system for each operator's voice characteristics. The system software has been written to allow speech characteristics for six operators to be stored in the computer memory at one time. If the Debug and Diagnostic sections of the program are deleted, the computer can store speech characteristics for eight operators. The basic approaches to speech recognition by machines especially as they apply to the VIP-100, are described in the following paragraphs, followed by a complete system description of the VIP-100 system as modified for this cartographic application.

B. Basic Approaches to Automatic Speech Recognition

Four processing functions are common to all automatic speech recognition systems. These functions as shown in Figure 1 consist of a microphone transducer, a preprocessor, feature extractor and a final decision level classifier. Early attempts at automatic speech recognition either deleted entirely the feature extraction process or utilized a simplified form of template matching. Experience with template matching soon led to the realization of its limitations. Slight variations of the individual speech samples of a particular word would result in gross misclassifications. This limitation resulted in the impractical requirement for a large memory containing a pattern and all its prototypes.

Considerable mathematical formalism has been developed for various automatic speech recognition processes. However, no general theory exists which can preselect the information bearing portions of the speech signal. Therefore, the design of the feature extractor is heuristic and must use ad hoc strategy. Only actual experimental data can determine the value of a particular feature set. It is this particular dilemma which has resulted in the recent increased emphasis given feature extraction research for pattern recognition systems.

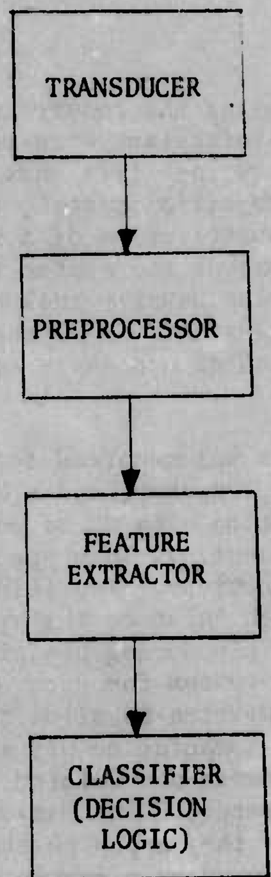


Figure 1. Pattern recognition process.

It is possible to form many transformations of the speech signal which would enhance certain properties and make them more easily detectable in an automatic speech-recognition system. However, speech is neither periodic nor aperiodic, but must be considered as a quasi-periodic signal so that analytical techniques that are developed must reflect temporal features of significance as well as spectral features. Maintaining this dual viewpoint throughout the analysis requires a modification of classical time-domain and frequency-domain analytical techniques. To retain both of these characteristics in a frequency analysis, a method which produces a short-duration spectrum is essential.

Frequency-domain representation of the speech signal is particularly advantageous since (1) it is known that the human auditory system performs a crude frequency analysis at the periphery of auditory sensation and (2) because it has been shown, by acoustical analysis of the vocalization system, that an exact description of the speech sounds can be obtained with a natural frequency concept model of speech production.

A periodic function of time possesses a power spectrum with finite amounts of power located a discrete points in the spectrum, commonly described as a line-spectrum. An aperiodic function that contains finite energy and is Fourier-transformable possesses an energy density spectrum that is a continuous function of frequency. For analyzing speech signals, it is desirable to obtain the spectral energy distribution and its variations as a function of time. Sufficient resolution must be maintained in both the frequency and time domains so that all of the information-bearing properties in both domains can be detected.

Spectrum analysis can be achieved either by direct analog circuitry or through the use of the Fast Fourier Transform (FFT) and a high speed digital computer. In both these methods, equivalent problems occur. The FFT produces a discrete spectrum which, with a sufficiently high sampling rate, approaches that of the continuous Fourier Transform. Many different types of data windows have been utilized in the FFT. The choice of the window is similar to the choice of the filter response in the analog spectrum analyzer. A "picket fence" effect can occur both in the FFT and the analog spectrum analyzer representing the contributions of the individual filters in the analog analyzer or the separate coefficients of the various terms in the FFT calculation. Analogous problems are introduced using linear predictive analysis in the selection of the number of coefficients employed in the process. In all cases, however, spectrum analysis is only the first step in the feature extraction process. Considerable additional processing is required in order to achieve the detection and recognition of the information-bearing elements (significant features) of the speech signal which has been transformed to accentuate these elements in the spectrum analysis process.

The final processing level after the recognition of the elemental speech units is the word decision logic. For isolated words, it is possible to examine the phonetic sequences produced by a feature extractor and to determine the closest match to a set of stored samples previously obtained for a given talker (or talkers). The decision involving the closest match is made at the end of the word and can be achieved with relatively simple processing techniques.

The VIP-100 equipment supplied to fulfill the requirements of this program, is a portable, isolated-word recognition system employing all of the processing functions previously described. Its specific implementation and carefully selected acoustic features result in a highly accurate and reliable equipment. Details of the techniques used in this equipment are included in the following paragraphs.

C. Description of the VIP-100

1. General

The VIP-100 is an automatic speech recognition machine designed for words spoken in isolation and can be automatically adapted for individual speakers and/or words. The system can be trained on-line and provides as an output a digital code which can be used to enter data into a computer, retrieve stored information, or control machine operations. The VIP-100 system consists of four basic units; they are the preprocessor, the minicomputer, the output interface and the Teletype. The preprocessor accepts the speech input from the microphone and converts it to logic signals which are then processed by the Nova 1200 minicomputer. The computer compares the input signal with stored references to determine which, if any, of the vocabulary words were spoken. If a correlation is found between the input speech and one of the vocabulary words, the appropriate ASCII message will be sent through the output interface If no correlation is found, an ASCII message indicating a REJECT condition will be transmitted through the interface. Figure 2 illustrates the ASCII code which is transmitted for each of the 15 vocabulary words plus the code for the REJECT condition.

Word	ASCII
0	060
1	061
2	062
3	063
4	064
5	065
6	066
7	067
8	068
9	069
ENTER	015
ERASE	010
CANCEL	030
MINUS	055
POINT	056
REJECT	007

Figure 2 ASCII characters, corresponding to 15 vocabulary words and the REJECT condition, which are transmitted by the VIP-100 to the digitizing computer.

Before an operator uses the VIP-100 in the speech recognition mode, the system is first optimized for both the vocabulary words selected and for the operator's particular manner of speech pronunciation by the use of a training routine. The operator speaks several utterances of each word in the vocabulary during the training operation. After training has been completed, the VIP-100 will be ready to recognize the chosen vocabulary words when they are spoken by the operator that trained the system. It is not necessary to retrain the system each time a particular operator uses the system. The operator training data may be stored in the active computer memory or on punched paper tape for use when needed. The appropriate tape with the stored data can be read into the system whenever the operator or vocabulary is changed. The system may be retrained for a single word, multiple words, or the complete vocabulary of 15 words at any time in order to accomodate vocabulary word substitutions or temporary changes in an operator's speech characteristics which may result from colds or other respiratory ailments.

The VIP-100 extracts the significant speech parameters (using hard-wired logic) necessary to characterize a particular word for a given speaker and stores these sample parameters such that they can be compared to new utterances and a word decision can be executed. The system includes a self-contained minicomputer in which word recognition is achieved by the application of predetermined decision algorithms. Such a system permits rapid "training" or adaption to new speakers and/or vocabularies. The system can be trained "on the spot" or can be externally programmed to insert speech characteristics previously obtained for the particular talker. Response time to the spoken words is virtually instantaneous; recognition outputs can be printed using the Teletype or visually observed on a display. Forced decisions can be made or "no decision" threshold criteria can be established, thereby requiring the speaker to repeat his utterance before a word decision is made. The system will operate to specifications in machine noise backgrounds as high as 85-90 dB.

2. Functional System Description

As described in Section II-B, a basic speech recognition system consists of four basic processing operations. Figure 3 is a block diagram of the VIP-100 recognition system showing a functional breakdown of the operations.

a. Transducer

The transducer employed in the system typically is a close-talking, noise-cancelling microphone mounted on a lightweight boom. The use of this type of microphone helps to reject background noise and provides speech signals of adequate fidelity. A Telex 1200 microphone mounted on a headband is used for this purpose.

b. Preprocessor

The preprocessor provides two principal functions. The first function is to shape the output from the microphone to remove irregularities and produce a normalized speech spectrum. The preamplifier associated with this operation provides 60 dB linear amplification and another 20 dB of limiting action for an overall processing range of 80 dB. The second function of the preprocessor is to perform a real-time spectral analysis of the equalized speech signal. The spectrum analyzer consists of a contiguous bank of 19 active bandpass filters ranging in center frequency from 260 Hz to 7626 Hz. These outputs are full-wave rectified and logarithmically compressed. This latter operation provides a 50 dB dynamic range and produces ratio measurements when subsequent features are derived from summation and differencing operations, thereby minimizing the input amplitude dependence. A detailed description of the principles used in the preprocessor is presented in the reference cited below.*

---

* T. B. Martin, "Acoustic Recognition of a Limited Vocabulary in Continuous Speech, PhD. Dissertation, University of Pennsylvania, May, 1970.

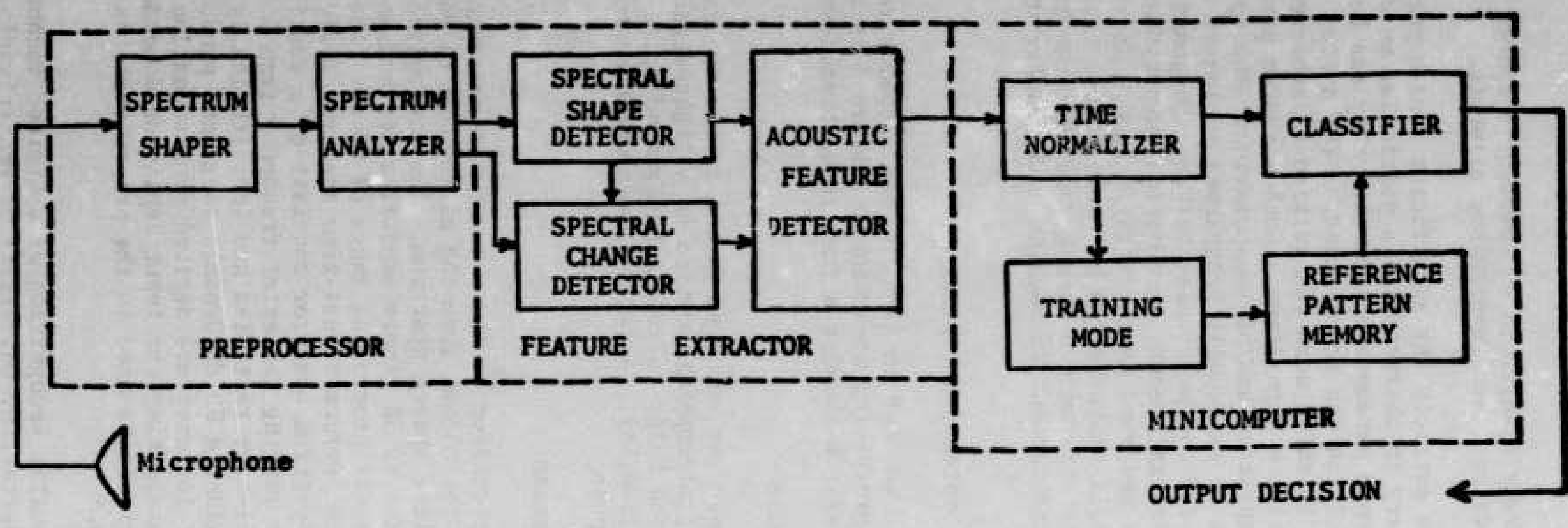


Figure 3 Block Diagram of VIP-100 Speech Recognition System

c. Feature Extractor

The key processing function in a pattern recognition system is the feature extractor. Although there are many acceptable classification techniques which can operate on a set of features (measurements), no classification scheme can compensate for an inadequate feature set.** The more optimum the feature set, generally the less complex the classifier need be to achieve a given accuracy. The various acoustic features used in the speech recognition system to be described have been tested extensively on very large speaker populations, in noisy as well as quiet backgrounds, and for many hours of on-line operations with untrained speakers. The features are useful for continuous speech applications as well as isolated word recognition systems. The selected feature set used for the VIP-100 is sufficiently general to make it possible to add new arbitrary words at any time to the system. The judicious selection and reliable extraction of these critical speech features distinguishes the VIP-100 from all other isolated-word speech recognition systems previously developed.

In the VIP-100 speech recognition system, the feature extraction process is accomplished principally by hard-wired logic. Using analog-threshold logic elements, various attributes of the speech signal are measured and significant speech parameters are extracted. The types of relevant acoustic features which can be extracted for speech signals are described in more detail in T. B. Martin's dissertation.

As shown in Figure 3, both the spectral shape and derivative are employed in the feature extraction process. The function of the spectral shape detector is to develop spectral derivative (dE/df) features indicating the overall spectrum shape. The spectral shape and its changes with time are continuously measured over the frequency range of interest. Combinations and sequences of these measurements are processed to produce a set of significant acoustic features.

The features used in the VIP-100 are a selected set (including complex combinations) of 32 acoustic features. Each feature is extracted by a combination of analog operations and binary logic. The output of the feature extractor consists of 32 binary signals, F_1, F_2, ... F_{32}.

The features are of two types, primary features and phonetic-event features. Features of the former category describe the spectrum directly by indicating local maxima and areas of increasing or decreasing energy with frequency (slopes). The latter category consists of features which represent measurements corresponding to phoneme-like events. Included in this set are vowels, nasals and fricatives.

Associated with the feature extractor process is the important requirement that accurate word boundaries be detected. The VIP-100 employs sophisticated pattern recognition techniques to accomplish this function. A hierarchy of features are measured and thresholds set to distinguish vocabulary words from background noise and extraneous speech utterances such as

** N. J. Nilsson, Learning Machines, New York, McGraw-Hill, 1965 coughs, sneezes, lip smacking, and breathing noises. The VIP-100 is remarkably immune to many of these types of disturbances. As a result, reliable word boundaries can be measured and used to accurately segment a word. This segmentation process is performed principally in hardware, although the final boundary detection is optimized through software in the minicomputer.

d. Classifier

The classification (or decision) process for the VIP-100 is performed in software using a minicomputer. Currently, a Nova 1200 16-bit minicomputer is used for this function. The minicomputer performs the multiplicity of functions shown in Figure 3. For a spoken word, the 32 encoded features and their time of occurrence are stored in a short term memory. When the end of the utterance is detected by the feature-extractor logic, the duration of the word is divided into 16 time segments and the features are reconstructed into a normalized time base. The pattern-matching logic subsequently compares these feature occurrence patterns to the stored reference patterns for the various vocabulary words and determines the "best fit" for a word decision. 512 bits of information (32 features mapped into 16 time segments) are required to store the feature map of an utterance or reference pattern.

3. Operation

a. System Considerations

The VIP-100 is an adaptive system which can be trained for individual talkers and/or words. Consequently, the system can be automatically adjusted or "tuned" to the voice characteristics of different users in a very short time period. By the inputting of a small number of training samples into the device to provide a reference set of features, the decision criteria for each word in the vocabulary can be modified or trained in an optimum manner. Thus, the system stores in memory an individual reference set of word features for each word in the vocabulary and for each talker in the system. Once system training is completed, new words spoken into the device during normal operation are compared with the stored references and a "closest fit" is selected as the recognized word. It is also possible to obtain a "no decision", or reject, when the characteristics of several words in the reference memory are very close to the spoken word. Since rejects may be permitted a predetermined percent of the time, a trade-off can be made between a reject (the speaker must repeat the word) and possible false responses. With this trade-off, it is possible to achieve high recognition accuracies and small substitution errors. The decision technique employed can be described most simply by briefly reviewing the operation of the system training and recognition mode.

b. Training Mode

During the training mode, the VIP-100 automatically extracts a time-normalized feature matrix for each repetition of a given word. A consistent matrix of feature occurrences (between repetitions) is required before the features are stored in the reference pattern memory. A template threshold factor is chosen such that a feature occurrence (in a given time segment) is considered valid only when it occurs a minimum number of times relative to the number of training samples. Usually, this threshold factor is set to be between 30-50% of feature occurrences within the training samples. An example of a reference feature matrix for the word "seven", based on 10 training samples and a threshold factor of 40%, is shown in Figure 4a. Figure 4b illustrates one training sample for this same word.

c. Recognition Mode

In the operational mode, each new word spoken into the system is processed in a manner analogus to the training procedure--i.e., features extracted, digitized and time normalized. The resultant test word matrix then is compared digitally to each stored reference matrix. Similarities and dissimilarities in each compared matrix are appropriately weighted and the net result provides a weighted correlation product. Correlation products also are generated after shifting the input word matrix $\pm$ 1 time segment. The stored reference word producing the highest overall correlation is selected as the test word.

In summary, the recognition system described provides high recognition accuracy because of the judicious choice of speech features measured during word utterances. The unique normalization and decision algorithms employed on the resultant feature sets permit tuning the system for individual talkers and produce extremely high recognition accuracies. The recognition process has been extensively tested in both hardware and software implementations and has been constructed in an economical method using integrated circuits.

4. Voice Input-Remote Control Unit

In most applications, it is desirable to physically locate the electronic equipment in a central location for ease of maintenance and logistics. To accomodate these applications the voice input to the system is achieved via a remote audio subsystem. In this mode of operation, more system flexibility is obtained by the use of a separate remote audio subsystem since few physical constraints are placed upon the location of a small remote input box. The remote audio subsystem, part of the Voice Input-Remote Control module, consists of microphone jack and equalizer circuitry, preemphasis circuitry and sufficient gain to transmit the audio to the remotely located preprocessor. A three position gain control is included in this subsystem, together with a Voice Level Meter to aid in correct input gain adjustments.

The Voice Input-Remote Control unit also houses thumbwheel switches which can be set to designate both the operator number for system usage and the word number for training purposes. When an operator comes on duty, he or she will select his or her user number which will then access the main memory file containing reference data for that person. These data would then be transferred to the active computer memory for use during operation. If the operator desires to train or retrain a word, he or she will select the appropriate word number and press a TRAIN button also located in the Voice Input-Remote Control unit. The speech preprocessor/recognition processor accepts this new training data and processes it such that the appropriate word reference data is stored in the minicomputer memory in place of the existing data for the word trained.

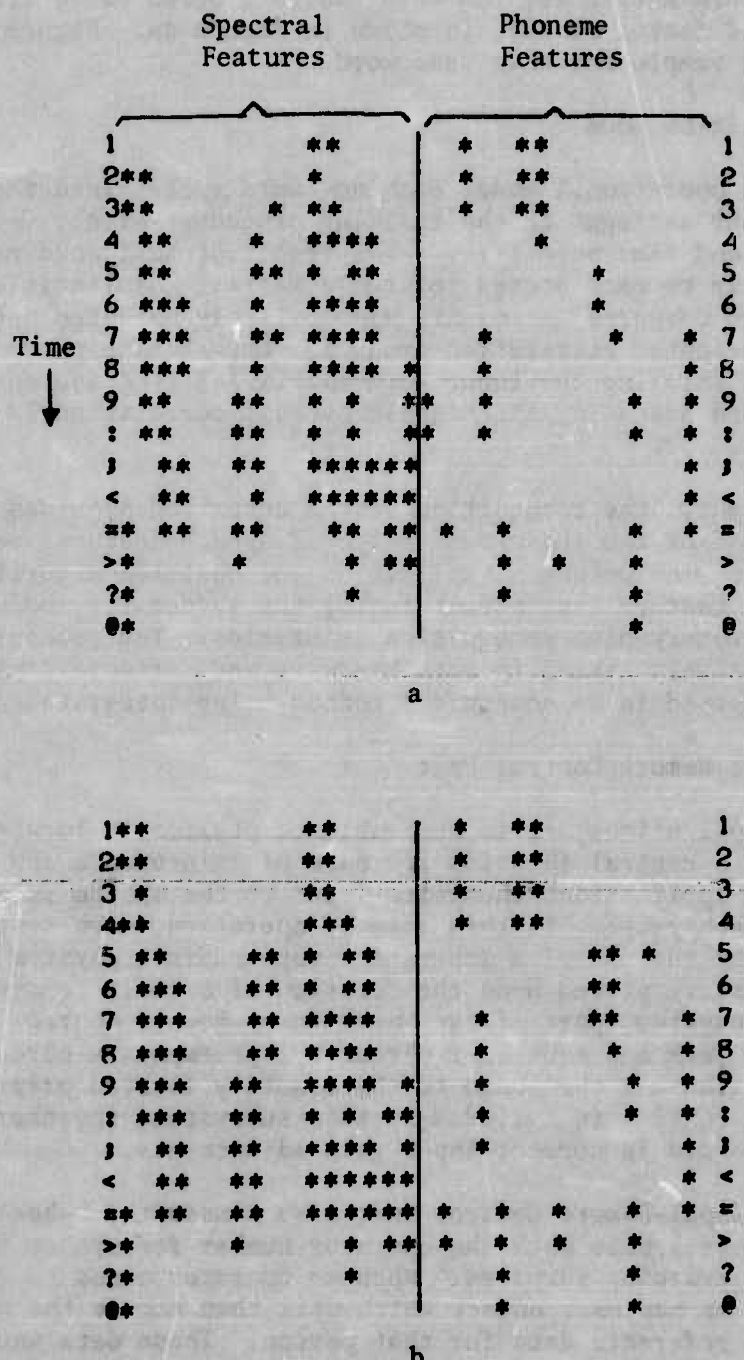


Figure 4. Reference feature matrix derived from 10 training repetitions of word "seven" (a), one training sample of this word (b)

5. Training Display

During normal cartographic data input by voice, a miniature display to be located on the cursor will be used by the operator to verify correct recognition of input words. This display will be provided by the digitizing equipment contractor. As an aid to the operator in the system training mode an auxiliary training display is provided. This display shows a symbol indicating the word being trained during the training mode and has additional functions which can aid the operator. The display module also includes a READY and a REJECT light and a REJECT SONALERT (an audible alert). During training, in addition to display of the word being trained, the display module by means of the READY light paces the operator. The REJECT SONALERT will sound once when the display changes during the training of the complete vocabulary.

During recognition this display is also operable. It will display four consecutive entries at one time. The REJECT and READY lights are also operable as is the REJECT SONALERT. The SONALERT provides, during normal recognition, an audible alarm indicating failure to recognize a word. It can be disabled by a switch to the rear of the module. The display unit is self contained with its own power supply and can be located where convenient.

6. Interface

To the Nova 1200 minicomputer, furnished GFE for inclusion in the VIP-100 system, two interface boards have been added. One of these boards is a standard Data General 4007/4010 Teletype interface which has been modified for operation at 2400 baud over a 20 ma current loop. This interface is used for transmission of ASCII characters (shown in Figure 2) representing recognized words to the digitizing computer from the VIP-100 system. This board occupies slot 5 in the Nova 1200 and is included in addition to the normal Teletype interface board located in slot 3 used to interface the control Teletypewriter. The connector for the slot 5 board is the same type as is used for the slot 3 board and has the same pin connections, therefore, care should be taken not to interchange the mating connectors to the digitizing computer and to the control Teletypewriter.

The other interface board added to the Nova 1200 is a Data General 4040 board with special TTI additional circuitry to allow the inputs to the computer from the proprocessor. This interface is located in slot 4 of the Nova 1200.

7. Software

The software is designed such that the system is interactive with the user and leads him through a set of possible operations. This procedure can be illustrated by showing some of the routine instructions presented to the user.

The operating software is provided in the form of punched paper tape which can be loaded into the Nova 1200 computer via Teletype or a high speed paper tape reader. Diagnostic software is provided to assist in checking the operation of both the speech preprocessor and special hardware associated with the recognition algorithm. The operating software has been written so that all control is performed via the Teletype, except for the functions of the Voice Input-Remote Control unit which is used to remotely train the system and to access operator reference data. The standard starting procedure for any system supplied with a programmers console* on the computer is to set the octal memory address 40 on the data switches first, then reset the computer with the RESET/STOP switch. The program can then be started by pressing the START switch. The system will respond by typing on the Teletype "TYPE 1 FOR INSTRUCTIONS". At this point the operator may type "1" followed by a carriage return to receive the following instructions:

```
CARTOGRAPHIC WORD RECOGNITION PROGRAM
TYPE:
I TO INPUT TRAINING DATA,
O TO OUTPUT TRAINING DATA,
A TO ASSIGN TRAINING PARAMETERS,
G TO GO TO RECOGNITION PHASE,
D TO GO TO DIAGNOSTIC PROGRAM,
? TO USE DEBUG1
```

The instruction printout can be skipped by typing one of the operating mode call characters listed above instead of a "1". All keyboard entries must be terminated with a carriage return before the computer will acknowledge the command. Incorrect keyboard entries may be cancelled (prior to pressing the return key) by pressing the rub-out key. The Teletype will respond with "?" to indicate it is again ready to accept an input command. Characters other than those specified above will be ignored and cause the Teletype to repeat its last message. The rub-out and message repeat features apply during all operational modes. Six of the seven possible operating modes may be entered either immediately before, or after the instruction type-out by pressing the appropriate key followed by a carriage return. The training mode is entered from the recognition mode as described below.

a. Recognition Mode

The recognition mode, entered by the use of the "G" command can be terminated at any time by depressing the CNTL key and the P key simultaneously on the Teletype keyboard. This action will cause the message "TYPE 1 FOR INSTRUCTIONS" to be typed and a new mode selection can then be made.

b. Training Mode

The training mode is selected by the use of controls at the Voice Input-Remote Control unit. The operator can train all words in the vocabulary by the following operations:

1) Dial appropriate SPEAKER NO. (1 through 6)
2) Dial 15 on WORD NO. switch**
3) Depress TRAIN indicator

* The GFE Nova 1200 includes a programmer's console.

** If the maximum vocabulary of 15 words is used. If a smaller vocabulary is used set the switch at the size being used.

A single vocabulary word can be trained simply by dialing the number of the word to be trained and repeating steps 1 and 3 above. The system can accomodate training data for six operators at a time.

The training routine is usually the first to be executed; its function is to adapt the recognition system for the voice characteristics of the particular user. For training the entire vocabulary, the training display will indicate the first word to be trained. Words being trained will not appear on the cursor display because no symbols are transmitted to the digitizing computer during training. The system is now ready to accept the specified number of training repetitions for each of the vocabulary words. Consecutive samples of a given vocabulary word are entered in sequence. That is, all samples of the first vocabulary word should be entered first. The training display will then indicate the second word to be trained and continue displaying that word until all training samples have been entered. The process will be continued until the entire vocabulary is trained. Remember to pause long enough for the READY light to reappear between each spoken word. The REJECT SONALERT will sound before the next word number appears on the display. When the training process is complete, the display will show a single "0" in the right-most position. The system will be ready to recognize spoken inputs and display the recognition outputs on both training and cursor displays. When the operator desires to retrain only a particular word the training display will indicate the word to be trained. When the correct number of samples have been entered into the system, the display will be cleared and the recognition mode takes over.

c. Input Training Data - I COMMAND

The system may be trained from a previously produced reference data paper tape by use of the "I" command. The reference data tape should be placed in the tape reader first; the reader control should then be set to the start position. The "I" command should then be entered on the keyboard followed by a carriage return. The computer will then ask for "SPEAKER NO?". After a number from 1 to 6 is entered followed by a RETURN, the paper tape will be read. The training data from the tape will replace the current training data (including vocabulary size) for the selected speaker. CAUTION, do not press any Teletype keys while the tape is being read.

d. Output Training Data - O COMMAND

The reference data compiled during training may be saved on punched paper tape for future use. The resulting tape will retrain the system for the particular operator and vocabulary when it is read into the system with the appropriate command. The reference data tape is produced with the output "O" command. Type the "O" command followed by a carriage return and turn the Teletype punch on. The computer will then ask for "SPEAKER NO?". After number 1 to 6 is entered followed by a RETURN, the computer will punch the paper tape. The reference training data will be punched out complete with leader at both ends of the tape. The Teletype will print "TYPE 1 FOR INSTRUCTIONS" when the tape is completed. Turn the punch off before entering an operating mode. The system will still be trained for the operator when the output routine is completed since execution of this routine does not modify the training data.

e. Assign Training Parameters - A COMMAND

The number of repetitions (from 1 to 10) used in training the vocabulary size (1 to 15) can be assigned by the A Command. After typing A-RETURN, the Teletype will respond with "NO OF REPS?". The operator should select the number of training repetitions desired. Optimum performance is obtained for 10 training repetitions. After the number of repetitions have been selected the Teletype will respond with "VOCABULARY SIZE?". A selection should then be made of the vocabulary size (normally 15). Remember if a vocabulary of fewer than 15 words is selected, that smaller vocabulary size is used for the train-all-words mode.

f. Diagnostic Programs - D AND ? COMMANDS

Two diagnostic software routines are provided to assist in system checkout. Both of these routines may be accessed in the same manner as the operating modes. The first diagnostic routine is designed to test the preprocessor and hardware interface to the computer. It is called by use of a "D" (Diagnostic) command. This routine will automatically test the bit counter and print "BIT COUNTER ERROR", "HIT CONTINUE TO TRY AGAIN" if an error is encountered. If errors are encountered on three successive passes, the bit counter circuit should be examined. The bit counter test (without errors) requires approximately 70 seconds running time. At the end of that time, if no errors are encountered, the Teletype will print "DISPLAY SHOULD READ 1248", "HIT ANY KEY TO BEGIN FEATURE TEST".

The feature test is conducted by the use of a special cassette tape supplied with the system. A tape cassette player, also supplied, should be connected to the tape input of the preprocessor and the input selector switch should be set to the TAPE position. Depress any key on the Teletype then play the cassette recording through the system. The computer will, after approximately 10 seconds of speech have been entered into the system, print a set of 40 numbers on the Teletype. This printout should be compared to the reference printout supplied with the system. A variation of more than 10% between this printout and the reference indicates that there may be a component failure within the preprocessor. The preprocessor circuitry for that feature then should be examined.

The second diagnostic routine is provided to assist in software debugging. The standard Data General Debug 1 program is available by entering the question mark (?) command. This program is described in detail in Data General Document Reference No. 093-000038-01.

g. Reloading the Program

If it becomes necessary at any time to reload the program into the Nova 1200 computer the following procedure should be followed.

(1) The Bootstrap Loader routine must be entered into the computer memory by the front panel switches. The loader is as follows (in octal representation):

Location	Data
17757	12644Ø
1776Ø	Ø6361Ø *
17761	ØØØ777
17762	Ø6Ø51Ø *
17763	1271ØØ
17764	1271ØØ
17765	1Ø7ØØ3
17766	ØØØ772
17767	ØØ14ØØ
1777Ø	Ø6Ø11Ø *
17771	ØØ4766
17772	Ø444Ø2
17773	ØØ4764

The above data words are to be used for loading from a Teletype. If a high speed reader is to be used change the right-hand "0" in the words indicated with * to a "2".

(2) Once the Bootstrap Loader is in memory the operator must load the Binary Loader tape into the reader, turn the reader on, set the computer data switches to 1777Ø and press RESET and then START. The Binary Loader program will then be read into memory.

(3) The recognition program tape is next loaded by a similar procedure to step 2 except that the switches should be set to 17777. Set the data switch Ø down, for reading from a Teletype, or up from reading from the high speed reader.

Section III

FINAL SYSTEM TESTS

A. Background of Test Data

Final testing of the cartographic word recognition system to establish performance levels was conducted by the use of tape recorded inputs from one female and 19 male talkers ranging in age from 16 to 50 years. Each of the 20 talkers recorded 360 test words and 150 training words. The training word sets consisted of 10 repetitions of each digit and each of the five control words. The test word sets consisted of 24 subsets of the complete vocabulary of 15 words. Test and training data were recorded in the same session. All recordings were made with a Telex model 1200 noise-cancelling microphone, one of two supplied with the system. Figure 5 is a near-field frequency response plot of the microphone used for the recordings. The other microphone supplied with the system was chosen as having a similar frequency response. All microphones used by TTI are measured by the use of a calibrated planewave tube.

The test data recordings described above were used as input data to the system for tests conducted at TTI before the system was shipped to RADC. The results of these tests which were verified by a representative of RADC are shown in Table I. Overall results of the test which included 7200 test words were: 99.375% correct responses, 0.347% incorrect responses, 0.246% rejects, and 0.041% no response. Figure 6 is an error matrix for the 20 talkers test. The "reject" category includes input words which the preprocessor could not identify because no correlation score was above a predetermined threshold. This threshold can be changed by the use of the Debug software feature described in Section III.C.7.f. However, no tests were conducted with a lowered threshold. Any appreciable decrease in the threshold can result in false recognition of extraneous noises. The "no response" category included words which were too short to exceed the minimum duration criterion established for the system. Only three no-response occurrences were noted, all by the same talker on the digit "8". The minimum duration criterion is not normally decreased because of the increased susceptibility of the system to respond to transient noise inputs.

The abbreviations used in Figure 6 for control words are as follows: En-ENTER, Er-ERASE, C-CANCEL, M-MINUS, P-POINT. These five control words are not necessarily optimum or final, but were chosen ad hoc by representatives of the digitizing computer contractors and TTI as a reasonable starting point. The flexibility of the VIP-100 based system allows the control words to be changed at any time.

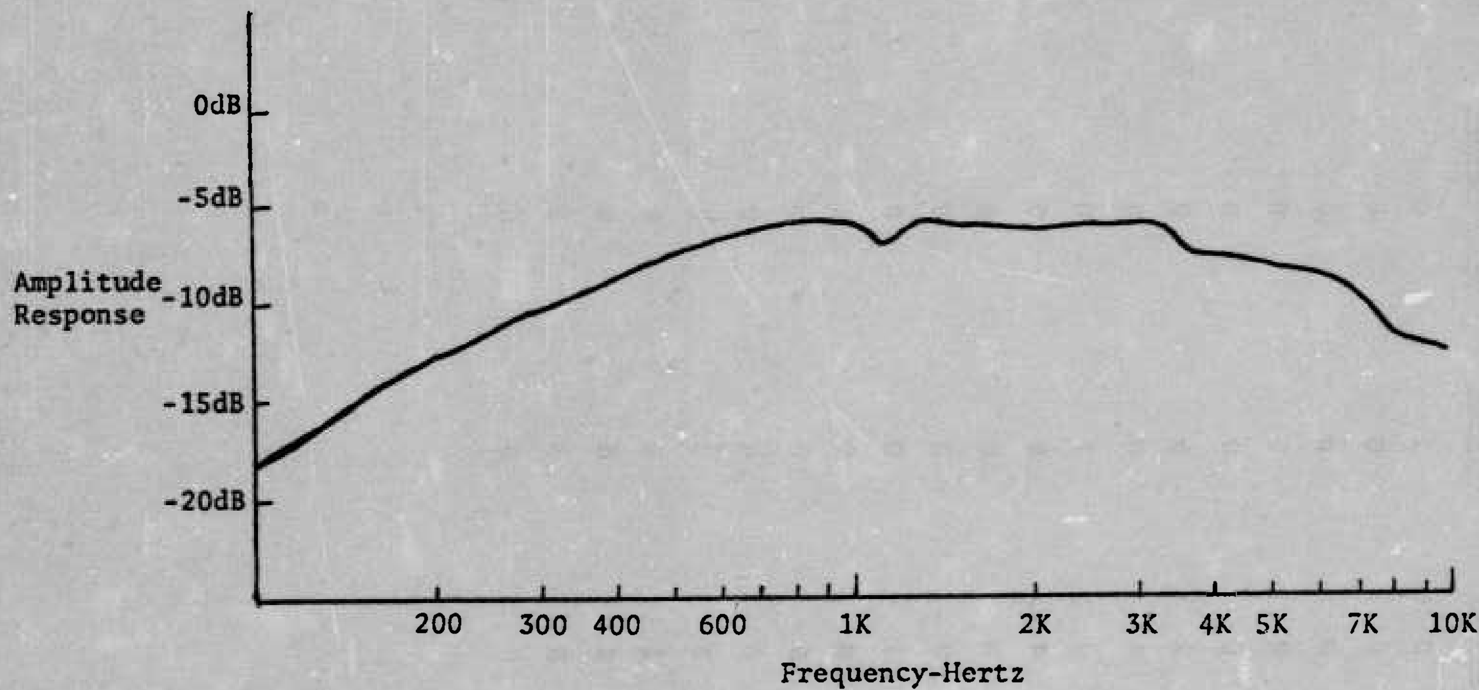


Figure 5 Measured near field frequency response of Telex Model 1200 microphone used for making data tape recordings. Microphone measured at 1/4" distance from output orifice of a calibrated Plane-Wave-Tube.

TABLE I RESULTS OF SYSTEM PERFORMANCE TESTS

Speaker	Errors	Rejects	No. Response	% Correct
PS	0	1	0	99.72
MW	0	0	0	100
MH	0	0	0	100
RP	0	5	0	98.6
RC	0	0	0	100
PL	1	6	0	98.06
KB	4	0	0	98.89
CC	2	1	0	99.16
JM	4	0	0	98.89
RB	0	0	0	100
EC	0	0	0	100
LS	1	0	0	99.72
MP	0	0	0	100
MS	5	0	0	98.6
TM	0	0	0	100
AT	2	0	0	99.44
AP	1	1	0	99.44
JG	0	0	0	100
JK	5	3	3	96.94
RV	0	0	0	100

Recognized \ Spoken	0	1	2	3	4	5	6	7	8	9	En	Er	C	M	P
0															
1						1									2
2															
3			2						1						
4															1
5															
6															
7															
8				1											
9						11									1
E_n				1											
E_r															
C															
M											1		1		
P						2									
Rej.					6						6	1			4
NR*									3						

Figure 6. Error matrix of 20 speakers each uttering 360 digits and control words. *NR denotes no response.

Section IV

CONCLUSIONS AND RECOMMENDATIONS

The major objectives of Alpha/Numeric Extraction Technique program have been achieved. A word recognition system suitable for use as an interface between a human operator and a cartographic digitizing computer has been developed from a standard VIP-100 word recognition system. The system will allow a cartographic operator to input by voice to the digitizer bathometric numbers from smooth sheets. The bathometric numbers are groups of from two to four digits indicating water depths. By the use of the voice input, the operator's hands are free to move the digitizer cursor over the smooth sheet which has been placed on a special digitizer table. The operator need not divert his attention from the table to a keyboard after noting each depth reading in order to input the data as has been the procedure heretofore. The use of voice input will greatly expedite data input and should enhance accuracy because the operator's hands are always free to move the cursor. Data should be inputted by voice at a rate of two to four times faster than by keyboard.

The word recognition system developed under this program should be a very useful tool in all phases of map making including, but not limited to, recording bathometric data. Extensive field tests with skilled operators under a variety of conditions will indicate how much of an increase in speed and accuracy can be expected. Such tests can also show how the system software can be expanded to accomplish data manipulation tasks in addition to word recognition.

APPENDIX

LOGIC EQUATIONS FOR PHONETIC-EVENT FEATURES

The recognition networks for phonetic-event features included in the feature set discussed in Section II, C.2.c, can be described by the use of logic equation as shown in Table II. These logic equations can be translated into equivalent logic diagrams. The notational rules for these logic equations are as follows:

1. An expression of the form $(\int_{T1} XQ1 - \int_{T2} YQ2)$ indicates that the excitatory quantity Q_2 and the inhibitory (subtractive) quantity Q_2 are integrated with time constants T_1 and T_2 and employ gain factors X and Y, respectively.

2. The analytical expression for the binary AND function will be of the form $C = A \cdot B$, where C represents the digital output of the AND gate for the two inputs A and B which can be in analog or digital form.

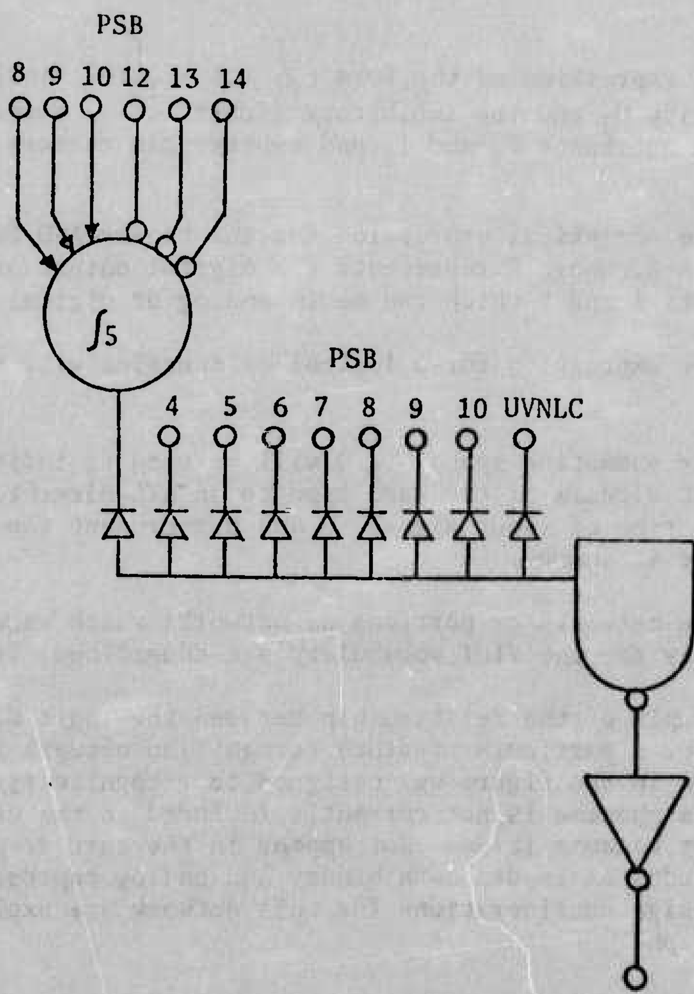
3. The expression for a logical OR function will be the form $C = A + B$.

4. The summation symbol $\sum_{m}^{n} Q$ will be used to indicate a plurality of (analog) input signals of the same type to an ATL element. In each case Q represents the type of input signal, m and n represent the interval over which the feature is summed.

5. The networks or portions of networks which were constructed or modified expressly for the VICI vocabulary are underlined with broken lines.

An example of the relationship between the logic diagram and the logic equation for a particular feature recognition network is shown in Fig. 7. The network shown in the figure was designed to recognize /ʃ/ in VIP-100 preprocessors. This phoneme is not currently included in the cartographic system feature array because it does not appear in the cartographic vocabulary. The network includes as inputs both binary and analog representation of positive slopes. Design considerations for this network are explained in the following paragraph.

This fricative consonant is characterized in wide-band speech by broad noise-like frequency bands above 1 kHz with a broad energy peak in the 2-3 kHz region. The resultant primary features useful for the detection of this characteristic are positive slopes (PSB) up through channels 10 or 11 of the VIP preprocessor. This phoneme is separated from the similar fricative, /s/ by the strength of positive slopes in channels 8 through 10 as compared with the positive slopes in channels 12 through 14. The phoneme /ʃ/ with a lower frequency concentration of energy in its spectrum as compared with /s/ can be expected to have stronger slopes in channels 8 through 10 as compared with channels 12 through 14. This separation is accomplished by the ATL element in the network as shown in the figure. The integration time constant associated with the inputs of the ATL element is 5 milliseconds. An input resistor value of 42.2K ohms is used for each input resulting in a gain factor of 0.8 times for each input. The unity gain input resistance for an ATL element is 34K ohms. Lower values of input resistance will therefore result in gains of greater than one. Binary representations of positive slopes in channels 4 through 10 together with the unvoiced noise-like (UVNLC) feature typical of fricatives are ANDed together with the output of the ATL element.



$$ʃ = \text{UVNLC} \cdot \text{PSB4} \cdot \text{PSB5} \cdot \text{PSB6} \cdot \text{PSB7} \cdot \text{PSB8} \cdot \text{PSB9} \cdot \left(\int_5 \sum_{8}^{10} \text{PSB} - \int_5 \sum_{12}^{14} \text{PSB} \right)$$

Figure 7. Logic diagram and equivalent logic equation for /ʃ/ recognition net word.

TABLE II PHONEME-LIKE FEATURE RECOGNITION LOGIC EQUATIONS (SHEET 1 of 3)

$$V/VL = \Big[(\int_5 \sum_1^3 1.7E - \int_5 \sum_4^5 1.16E) + (\int_5 \sum_1^3 1.7E - \int_5 \sum_{10}^{12} 1.7E) \Big] \cdot 1.27 \Big[(\int_5 \sum_1^4 1.7E - \int_5 \sum_{15}^{19} 1.7E) +$$

$$(\int_5 \sum_7^{11} 1.7E - \int_5 \sum_{16}^{19} 1.7E) - \int_9 (\int_5 \sum_{15}^{19} 1.4E - \int_5 \sum_{10}^{14} 1.4E) \Big] \cdot 1.27 \Big[(\int_5 \sum_{11}^{14} 1.7E - \int_5 \sum_{15}^{18} 1.7E) +$$

$$(\int_5 \sum_6^9 1.7E - \int_5 \sum_{10}^{13} 1.7) + (\int_5 \sum_1^4 1.7E - \int_5 \sum_8^{11} 1.7E) - \int_9 (\int_5 \sum_{15}^{19} 1.4E - \int_5 \sum_{10}^{14} 1.4E) \Big]$$

25

$$BV = \Big[(\int_5 \sum_6^8 2.75E - \int_5 \sum_9^{11} 2.75E) + (\int_5 \sum_7^9 2.75 - \int_5 \sum_{10}^{12} 2.75E) + NSB8 \Big]$$

$$\varepsilon_2 = V/VL \cdot MAX3 \cdot (MAX9 + MAX10) \cdot NSB4$$

$$3 = V/VL \cdot BV \cdot NSB9 \cdot NSB10 \cdot NSB11 \cdot (NSB8 + NSB12) \cdot (1.4 \sum_8^9 NSB - 1.4 \sum_6^7 NSB)$$

$$r = NSB3 \cdot NSB9 \cdot NSB10 \cdot NSB11 \cdot (NSB8 + NSB12) \cdot (1.4 \sum_8^9 NSB - 1.4 \sum_6^7 NSB) \cdot$$

$$(1.4 \sum_3^4 NSB - 1.4 \sum_6^7 NSB) \cdot (\int_5 \sum_7^9 2.75E - \int_5 \sum_{10}^{12} 2.75E) \cdot (1.15 \sum_8^{10} NSB - 1.15 \sum_2^4 NSB)$$

TABLE II PHONEME-LIKE FEATURE RECOGNITION LOGIC EQUATIONS (SHEET 2 of 3)

$$w = V/VL \cdot BV \cdot \Big[\big(\int_5 \sum_1^{3} 1.15E - \int_5 \sum_7^{10} 1.15E\big) \cdot \big(\int_5 \sum_4^{6} 1.15\ NSB - \int_5 \sum_8^{11} 1.15\ NSB\big) +$$

$$\big(\int_5 \sum_1^{3} 1.15E - \int_5 \sum_5^{9} 1.15E\big) \cdot \big(\int_5 \sum_3^{5} 1.15\ NSB - \int_5 \sum_8^{11} 1.15\ NSB\big) \Big]$$

$$UVNLC = \overline{V/VL} \Big[\big(\int_{14} \sum_{15}^{19} 1.4E - \int_{14} \sum_{10}^{14} 1.4E\big) + \big(\int_{14} \sum_{15}^{19} 1.4E - \int_{14} \sum_5^{9} 1.4E\big) \Big] \cdot \Big[\underline{BRST} \Big] \cdot$$

$$\Big[\underline{\big(\int_{14} \sum_{15}^{19} 1.4E - \int_{14} \sum_1^{4} 1.7E\big)} \Big]$$

$$s = UVNLC \Big[PSB\ (5\cdot 6\cdot 7\cdot 8\cdot 9\cdot 10\cdot 11\cdot 12\cdot 13) + PSB\ (6\cdot 7\cdot 8\cdot 9\cdot 10\cdot 11\cdot 12\cdot 13\cdot 14) +$$

$$PSB\ (7\cdot 8\cdot 9\cdot 10\cdot 11\cdot 12\cdot 14\cdot 15) \Big]$$

$$\Lambda = V/VL \cdot BV \cdot \Big[\big(1.4 \sum_6^{7} PSB - 1.4 \sum_8^{9} PSB\big) + \big(1.4 \sum_8^{9} NSB - 1.4 \sum_6^{7} NSB\big) \Big] \cdot$$

$$\Big[PSB\ (2+3) + PSB\ (3+4) \cdot PSB2 \cdot \big(\int_5 \sum_{10}^{13} 2.75E - \sum_{16}^{19} 2.75E\big) \Big]$$

TABLE II PHONEME-LIKE FEATURE RECOGNITION LOGIC EQUATIONS (SHEET 3 of 3)

$$n = \left[NSB\ (1 \cdot 2 \cdot 3) \cdot EG_1 \cdot EG_2 \right] + \left[EG_1 \cdot V/VL \cdot NSB\ (2 \cdot 3 \cdot 4) \cdot \underline{BRST} \cdot \left(\int_5 \sum_1^4 1.7E - \int_5 \sum_6^9 1.7E \right) \right.$$

$$\left. \cdot \left(\int_{2.2} \sum_1^4 1.15\ NSB - \int_{2.2} \sum_1^4 1.15\ NSB - \int_{2.2} \sum_6^9 1.15\ NSB - \int_{2.2} \sum_8^9 1.15\ PSB \right) \right]$$

$$\text{Energy Gap} = EGap_1 + EGap_2 = \left[\int_{2.2} 3.4 \int_{55} \left(\int_5 \sum_5^9 .25E - \underline{5.63\ UVNLC} \right) - \int_{2.2} 3.4 \int_5 \sum_5^9 .25E \right) \right] +$$

$$\left[\left(\int_{2.2} 3.4 \int_{35} \left(\int_5 \sum_{10}^{14} .25E - \underline{5.63\ UVNLC} \right) - \int_{2.2} 3.4 \int_5 \sum_{10}^{14} .25E \right) \right]$$

☆U.S. GOVERNMENT PRINTING OFFICE: 1975-614-357/58

MISSION
of
Rome Air Development Center

RADC is the principal AFSC organization charged with planning and executing the USAF exploratory and advanced development programs for information sciences, intelligence, command, control and communications technology, products and services oriented to the needs of the USAF. Primary RADC mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, and electronic reliability, maintainability and compatibility. RADC has mission responsibility as assigned by AFSC for demonstration and acquisition of selected subsystems and systems in the intelligence, mapping, charting, command, control and communications areas.

AMERICAN REVOLUTION BICENTENNIAL 1776-1976